Comparison of Survey Estimates of the Finite Population Variance

Jean-Yves P. COURBOIS and N. Scott URQUHART

The Environmental Monitoring and Assessment Program (EMAP) of the U.S. Environmental Protection Agency has conducted several probability surveys of aquatic resources. Such surveys usually have unequal probability of including population elements in the sample. The Northeast lakes survey, which motivated this study of variance estimation, was such a survey. We examine ten estimators for the finite population variance using a Monte Carlo factorial experiment that considers three population characteristics. The results show that the correlation between the inclusion probabilities and the response is the most important factor that differentiates the estimators. Under conditions of low correlation (approximately < 0.4), a common feature in environmental surveys, the sample variance is best, elsewhere, two ratio estimators, one based on consistency and the Horvitz-Thompson Theorem (HT) and the other based on the Yates-Grundy form, behave similarly and best.

Key Words: Finite population sampling; Horvitz-Thompson estimation; Unequal probability sampling.

1. INTRODUCTION

Objectives of a survey sometimes include estimating a finite population variance. For instance, Thompson (1992, p. 33) described a method for sample size determination which requires an estimate of the population variance. This estimate ideally comes from a pilot survey. The U.S. Environmental Protection Agency's Environmental Monitoring and Assessment Program-Surface Waters Northeast Lakes Pilot study (EMAP-lakes) has this characteristic (Larsen and Christie 1993), as it aims to gain population insight toward planning future sampling strategies. Toward this goal, in addition to others, the data from that pilot project will be used to estimate population variances (Larsen, Thorton, Urquhart, and Paulsen 1995; Urquhart, Paulsen, and Larsen 1998). In other situations the population variance may be a parameter of interest in its own right. For example, in ecological applica-

Jean-Yves P. Courbois is Postdoctoral Research Associate, University of Washington, Statistics Department Box 354322, Seattle, WA. 98195-4322 (E-mail: pip@stat.washington.edu). N. Scott Urquhart is Senior Research Scientist, Department of Statistics, Colorado State University.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 9, Number 2, Pages 236–251
DOI: 10.1198/1085711043596

tions a small population variance indicates a uniform population. Sound estimators of the population variance will prove useful.

The sample variance is regarded as the estimator of choice in the case of simple random sampling (Cochran 1977; Thompson 1992); however, the EMAP-lakes pilot project uses a complex sample design to select a sample of lakes (Larsen et al. 1995). A complex sample design assigns unequal inclusion probabilities to population elements.

Sampling with unequal inclusion probabilities is common in multipurpose environmental surveys. The U.S. Environmental Protection Agency's National Stream Survey (NSS) (Stehman and Overton 1994a), the multipurpose monitoring plan for the California Bight (Stevens 1994), and the EMAP-lakes pilot project (Larsen and Christie 1993) all use complex designs that arise because of multiple survey objectives and practical constraints. For a complex design the sample variance is not design unbiased. How should we estimate a finite population variance with a complex design?

Särndal, Swensson, and Wretman (1992, sec. 5.9) suggested three estimators for this case but did little to compare them stating, simply, that the estimators should not behave differently. The first, however, is a ratio estimator but the other two are not. Stehman and Overton (1994b) recommended a ratio estimator as well. Other than consistency, they provided no properties either. Liu and Thompson (1983) proved that an estimator based on the Yates-Grundy form is admissible among unbiased estimators and that a "generalized" Horvitz-Thompson estimator is inadmissible. Taken together, these recommendations provide only unsubstantiated suggestions, not a solution to the problem of which estimator should be used. Our sampling literature lacks a Monte Carlo study of this estimation problem (Royall and Cumberland 1981; Stehman and Overton 1994a), which can provide pragmatic answers for situations where more than one estimator exists but an analytical comparison proves difficult, if not intractable.

We examine ten different estimators for the finite population variance. The first three are nonratio forms of the estimators suggested by the literature above, the next six are the ratio forms of these estimators, and the last is a naive estimator: the sample variance.

The sample variance interests us because of its simplicity and how likely it is to be used. For this estimator, we ask: "When is it applicable?" and "What can go wrong if it is naively calculated from a sample collected with a complex design?" As for the rest of the estimators, we ask "Which of the ten estimators performs better than the others and under what population and sample design conditions?"

Our results rely on a simulation study based on both real and artificial populations. The real populations cover a range of possible populations practitioners might encounter. The artificial populations, on the other hand, provide control over simulation parameters. We use a population space approach by varying simulation parameters in a factorial structure to make inference to an encompassed set of possible parameters (Stehman and Overton 1994a), varying three parameters: the correlation between response and inclusion probabilities; the variance of the inclusion probabilities; and the sample size. If a practitioner can place their population and sample design into one of our categories and choose an estimator based on our results, then we have succeeded. In other cases, practitioners should perform similar simulations.

We restrict our comparisons to probability-proportional-to-size designs that have a simple method of sample selection. Although thus restricted, our results should provide guidance in many other situations.

Section 2 introduces the notation and derives the estimators and their characteristics. Section 3 describes the empirical experiment we use to compare the ten estimators. Section 4 presents the results of the experiment: The sample variance is extremely resistant to the design and hence can be used under general conditions; in other cases, a ratio form of the HT estimator performs best. Section 5 discusses the importance of these results.

2. THE ESTIMATORS

Let $\mathcal{U}$ be the universe of $1 < N < \infty$ elements and y_i be a response variable that is measurable without error on every unit $i \in \mathcal{U}$. A sample design, $p(S)$, assigns a probability of selection to every possible sample, $S \subseteq \mathcal{U}$, and thus first-order inclusion probabilities to each unit $i \in \mathcal{U}$, π_i, and second-order inclusion probabilities to all pairs of units $i, j \in \mathcal{U}$, π_{ij}. Finally, define Z_i as the sample membership indicator for element i, $Z_i = 1$ if $i \in S$ and $Z_i = 0$ if $i \notin S$.

The population variance can be represented in three ways, each of which motivates a set of three estimators. The first form provides basis for method of moment estimators,

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \overline{y})^2 ;$$

the second demonstrates that the variance is the sum of two "averages," the average of the squared responses and the square of the average, and leads to the Horvitz-Thompson estimators

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} y_i^2 - \frac{N}{N-1} \overline{y}^2; \qquad (2.1)$$

and the third form is due to Yates and Grundy (1953), who derived a similar form for the variance of the HT estimator of the population total

$$\sigma^2 = \frac{1}{2N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} (y_i - y_j)^2.$$

2.1 The Design-Based Estimators

The Horvitz-Thompson theorem and a quadratic form of it (Corollary 1) suggest three design-based estimators. The first is an unbiased method-of-moment (MOM) estimator,

$$S_{\Pi}^2 = \frac{1}{N-1} \left(\sum_{i=1}^{N} y_i^2 \frac{Z_i}{\pi_i} - \frac{1}{N} \sum_{i=1}^{N} \sum_{i=1}^{N} y_i y_j \frac{Z_i Z_j}{\pi_{ij}} \right).$$

Table 1. Descriptive Properties of the Estimators

		y-invariance		
Estimator	*Unbiased*	*Scale*	*Location*	*≥ 0*
s^2	No	Yes	Yes	Yes
S^2_Π	Yes	Yes	No[a]	No
$\hat{S}^2_\Pi$	No	Yes	No[a]	?
$\tilde{S}^2_\Pi$	No	Yes	No[a]	No
S^2_π	No	Yes	No	No
$\hat{S}^2_\pi$	No	Yes	Yes	Yes
$\tilde{S}^2_\pi$	No	Yes	No	No
S^2_*	Yes	Yes	Yes	Yes
$\hat{S}^2_*$	No	Yes	Yes	Yes
$\tilde{S}^2_*$	No	Yes	Yes	Yes

NOTE: [a] These estimators are *y*-location invariant in expected value, see Section 2.

This estimator is y-scale invariant (Table 1), but is location invariant only in expected value, for any constant c,

$$S^2_\Pi(y+c) = S^2_\Pi(y) + \frac{2c}{N-1}\left(\sum_{i=1}^{N}\frac{y_i Z_i}{\pi_i} - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{y_i Z_i Z_j}{\pi_{ij}}\right) + \frac{c^2}{N-1}\left(\sum_{i=1}^{N}\frac{Z_i}{\pi_i} - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{Z_i Z_j}{\pi_{ij}}\right).$$

The two additional terms on the right-hand side compare the traditional Horvitz-Thompson estimator with a generalized Horvitz-Thompson estimator (Corollary 1) of the population total and size, respectively. Both terms have expectation zero. This estimator can assume negative values.

The next design-based estimator stems from the bilinear form of the variance (2.1). Use of the Horvitz-Thompson estimators, $\sum_{i=1}^{N} y_i^2\, Z_i/\pi_i$ and $(\sum_{i=1}^{N} y_i\, Z_i/\pi_i)^2$, to estimate the two sums results in the HT estimator

$$S^2_\pi = \frac{1}{N-1}\left(\sum_{i=1}^{N} y_i^2 \frac{Z_i}{\pi_i} - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \frac{Z_i Z_j}{\pi_i \pi_j}\right). \qquad (2.2)$$

This estimator is biased, not y-location invariant, and can assume negative values (Table 1). However, the bias of this estimator is likely to be small,

$$E(S^2_\pi) = \sigma^2 - \frac{1}{N}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j - \sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \frac{\pi_{ij}}{\pi_i \pi_j}\right).$$

Finally, the “batch” approach to HT estimation, Corollary 2, and the third form of the population variance suggest the unbiased YG (Yates-Grundy) estimator

$$S_*^2 = \frac{1}{2N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{Z_i Z_j \, (y_i - y_j)^2}{\pi_{ij}}.$$

This estimator is unbiased, location and scale invariant, and strictly positive (Table 1). Liu and Thompson (1983) found this estimator to be admissible in the set of unbiased estimators.

2.2 Ratio Forms of the Design-Based Estimators

The survey sampling literature suggests that, when estimating a population mean under a complex design, the HT estimator of the population size should be used in the denominator in place of the actual population size (Särndal et al. 1992; Thompson 1992), that is, one should estimate N with $\hat{N} = \sum_i^N Z_i/\pi_i$. Our estimators that have $\hat{N}(\hat{N}-1)$ in their denominator are denoted with a hat such as $\hat{S}_\Pi^2$ and abbreviated with a “-R”, indicating a “ratio” estimator, such as the MOM-R estimator. However an alternative estimator of N^2 is $\widetilde{N^2} = \sum\sum_{i,j=1}^N Z_i Z_j/\pi_{ij}$. The estimators that have $\widetilde{N^2} - \hat{N}$ are denoted with a tilde and abbreviated with a “-GR”, generalized ratio estimator.

The ratio forms of the estimators do not share the same properties as their unweighted forms (Table 1). Note that estimating population size is often necessary in environmental sampling when N is unknown, frequently because the sampling frame is imperfect, a phenomenon experienced in the EMAP survey of Northeast lakes.

2.3 Second-Order Inclusion Probability Approximations

The MOM, YG, and all the GR-estimators require second-order inclusion probabilities. Because the exact computation of second-order inclusion probabilities is cumbersome under the design described earlier (Hidiriglou and Gray 1980), we use an approximation suggested by Stehman and Overton (1989): $\pi_{ij} = \frac{(n-1)\pi_i\pi_j}{2n-\pi_i-\pi_j}$, with π_i substituted for π_{ii}. We also investigated the approximation suggested by Hartley and Rao (1962) and found it made no difference in our variance estimates.

2.4 Analytical Comparisons of the Design-Based Estimators

The HT-based estimators estimate the two parts of the variance in different ways:

$$\begin{array}{lrcl} S_\pi^2 \propto & \sum_{i=1}^N y_i^2 \frac{Z_i}{\pi_i} & - & \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^N y_i y_j \frac{Z_i Z_j}{\pi_i \pi_j}, \\ S_\Pi^2 \propto & \sum_{i=1}^N y_i^2 \frac{Z_i}{\pi_i} & - & \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^N y_i y_j \frac{Z_i Z_j}{\pi_{ij}}, \\ S_*^2 \propto & \sum_{i=1}^N y_i^2 \sum_{j=1}^N \frac{Z_i Z_j}{\pi_{ij}} & - & \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^N y_i y_j \frac{Z_i Z_j}{\pi_{ij}}. \end{array} \qquad (2.3)$$

Each estimator either uses the HT estimator or the estimator suggested by Corollary 2 for a part of the variance. The HT estimator, S_π^2, uses only the HT theorem, the YG estimator, only Corollary 2, and the MOM estimator, S_Π^2, a mixture of the two.

Examination of the estimators reveals that if we consider the HT Theorem as a basis for finite population estimation, the MOM estimator is the most "natural." It estimates the first average with the HT estimator and the second average with the batch estimator.

2.5 A Naive Estimator

For reasons previously noted, we are interested in the consequences of using the unweighted sample variance to estimate σ^2 under complex designs. We call this the "naive" estimator because it ignores the sampling probabilities: $s^2 = \sum_i^n (y_i - \overline{y})^2/(n-1)$, where $\overline{y}$ is the sample mean.

The naive estimator has many good properties (Table 1), although it is biased:

$$E\left(s^2 - \sigma^2\right) = \frac{1}{n}\sum_{i=1}^{N} y_i^2 \left(\pi_i - \frac{n}{N}\right) - \frac{1}{n(n-1)}\sum_{j=1}^{N}\sum_{j\neq i}^{N} y_i y_j \left(\pi_{ij} - \frac{n(n-1)}{N(N-1)}\right). \tag{2.4}$$

This bias decreases as the variance of the π_i in the population decreases. Also, because the terms $\pi_i - n/N$ and $\pi_{ij} - n(n-1)/N(N-1)$ each sum to zero, the bias decreases when the y and π are uncorrelated. These two criteria form the basis of our simulations.

3. THE SIMULATION EXPERIMENTS

Our simulation experiments address two questions: (1) Under what conditions does the sample variance perform well? and (2) Which of the estimators outperform the others and under what conditions? The simulations use both real and artificial populations. The real populations come from two social and two environmental surveys. We created the artificial populations by varying three population parameters that describe the joint distribution of the inclusion probabilities and the response.

For each simulation, we selected 10,000 samples and for each sample calculated the value of the variance estimators. All simulations were programmed and run in SAS-AML$^{(\mathrm{TM})}$; copies of code are available from the first author.

Samples were selected by constructing a line whereon each universe element is represented by a line segment whose length equals the element's inclusion probability in random order. After a random start, a systematic sample was selected by taking equal sized steps along the line and including in the sample every element on whose line segment a step lands (Stevens 1994).

3.1 The Real Populations

Seventeen datasets from four surveys serve as the real populations (Table 2). The United States Department of Health and Human Service's 1995 National Nursing-Home Survey (NNHS) provides the first universe, a sample of 1,368 nursing homes. The response is the

Table 2. The Real Populations. Under *source*, NNHS refers to the National Nursing Home Survey, SSW refers to the Swedish municipality population, NSS refers to the National Stream Survey, and EMAP refers to the EMAP-lakes pilot survey (Section 3.1). ANC stands for Acid neutralizing capacity for both the streams and the lakes. The parameter γ measures the distance a sample design is from simple random sampling (Section 3.2).

	Source	*x*	*y*	*cv (x)*	γ	ρ	*N*
1	NNHS	inverse sample weight	# of employees	0.70	0.02	0.83	1368
2	SSW	1975 population	# of employees, 1984	1.83	0.22	0.97	284
3	SSW	”	# of seats in council	1.83	0.22	0.68	”
4	NSS	watershed area	ANC	1.38	0.06	−0.15	1630
5	NSS	”	stream depth	1.38	0.06	0.17	”
6	EMAP	watershed area	ANC	5.30	0.57	0.13	490
7	”	sqrt. watershed area	”	1.64	0.24	0.13	”
8	”	log. watershed area	”	0.33	0.01	0.04	”
9	”	lake area	”	2.65	0.48	−0.06	”
10	”	sqrt. lake area	”	1.09	0.13	−0.09	”
11	”	log. lake area	”	0.49	0.03	−0.11	”
12	”	watershed area	Secci depth	5.08	0.57	−0.04	434
13	”	sqrt. watershed area	”	1.61	0.26	−0.04	”
14	”	log. watershed area	”	0.33	0.01	−0.03	”
15	”	lake area	”	2.50	0.48	0.11	”
16	”	sqrt. lake area	”	1.05	0.14	0.19	”
17	”	log. lake area	”	0.47	0.03	0.22	”

number of employees and the auxiliary variable is the inverse of the sample weight (Table 2 and Figure 1).

Särndal et al. (1992) supplied the second universe: data from the 284 municipalities of Sweden. The number of municipal employees in 1984 and the total number of seats on the municipal council serve as the responses, while for both of these the 1975 municipality population is the auxiliary variable (Table 2 and Figure 1) .

USEPA’s national stream survey (NSS) (Mitch et al. 1990) provides a fourth and a fifth population on a universe of 1,630 stream traces (Table 2 and Figure 1). The responses are the acid neutralizing capacity (ANC) and the stream depth. The auxiliary variable is the watershed area.

The final eleven populations come from the EMAP-lakes pilot project (Larsen and Christie 1993). From this dataset, we take two responses, ANC and Secci depth (Secci depth is a measure of the transmission of visible light into a body of water), and construct six different auxiliary variables from lake characteristics: the lake’s watershed area, the lake area, and each of these after the square-root and natural logarithm transformations (Table 2 and Figure 2).

3.2 The Artificial Populations

Section 2.5 demonstrated that two factors affect the magnitude of the bias of the naive estimator (Equation (2.4)); first, the estimator’s bias depends on the correlation between the response and the inclusion probabilities, and second by how “far” the design is from simple random sampling. These parameters serve as the bases for our experiment. We build

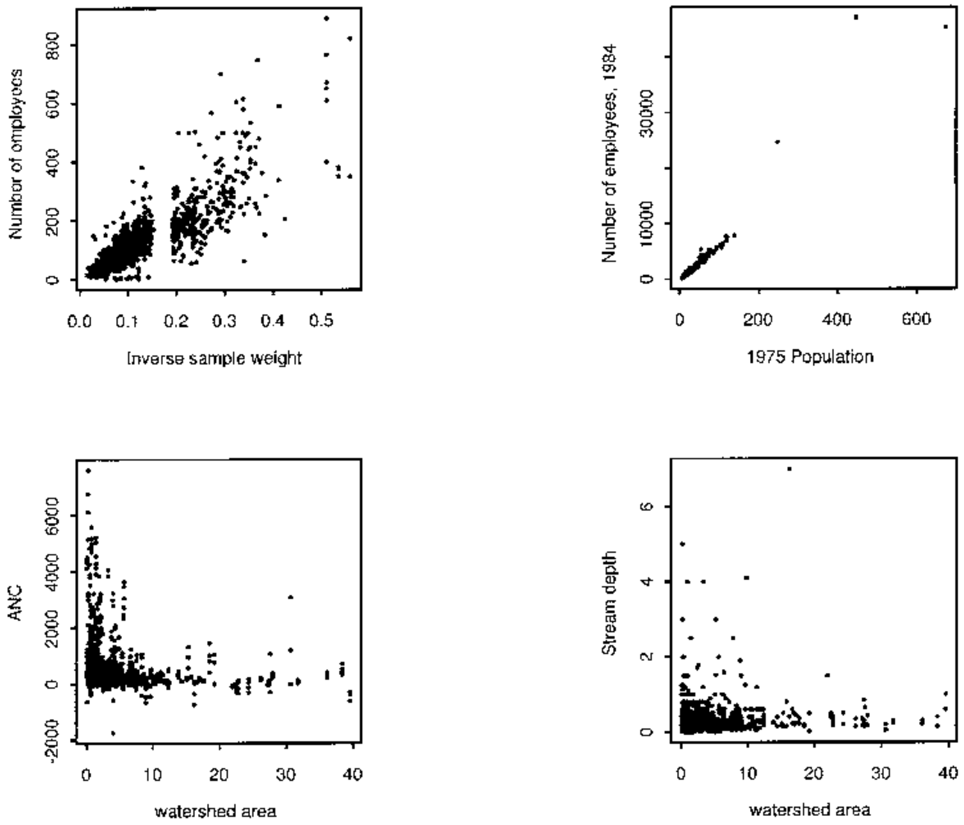

Figure 1. The auxiliary variable and response for real populations 1, 2, 3, 4 (Table 2 and Section 3.1).

populations with correlations that range from -0.95 to 0.95. In the EMAP-lakes data, the relationships between the chemical responses and the measurements of lake size have correlations from -0.30 to 0.27, (Figure 2). The variance of the inclusion probabilities, $\sigma_\pi^2 = \sum_{i=1}^N (\pi_i - \overline{\pi})^2/(N-1)$, measures the distance the design is from simple random sampling (McDonald 1996). This variance is constrained by the inequality

$$0 \leq \sigma_\pi^2 \leq \frac{n(N-n)}{N(N-1)}. \tag{3.1}$$

When $\sigma_\pi^2 = 0$ the design is simple random sampling; however, as $\sigma_\pi^2 \to \frac{n(N-n)}{N(N-1)}$ the design degenerates to one that assigns probability 1 to the n units with the largest auxiliary variable values and 0 to all other units. Our simulations use four levels of the ratio, $\gamma = \sigma_\pi^2 / \frac{n(N-n)}{N(N-1)}$, $(0.1, 0.2, 0.3, \text{ and } 0.4)$, a range which covers practical applications.

For each of these 68 simulations, 17 levels of correlation by 4 levels of γ, an auxiliary variable, $x_i,\ \ i = 1, \ldots, N$, was a draw from a one parameter gamma distribution. This distribution describes positively skewed populations such as what is found in nature (e.g., lake size). These were standardized to have unit variance. After the x_i were generated, the responses, $y_i,\ \ i = 1, \ldots, N$, were generated using a linear regression model: $y_i =$

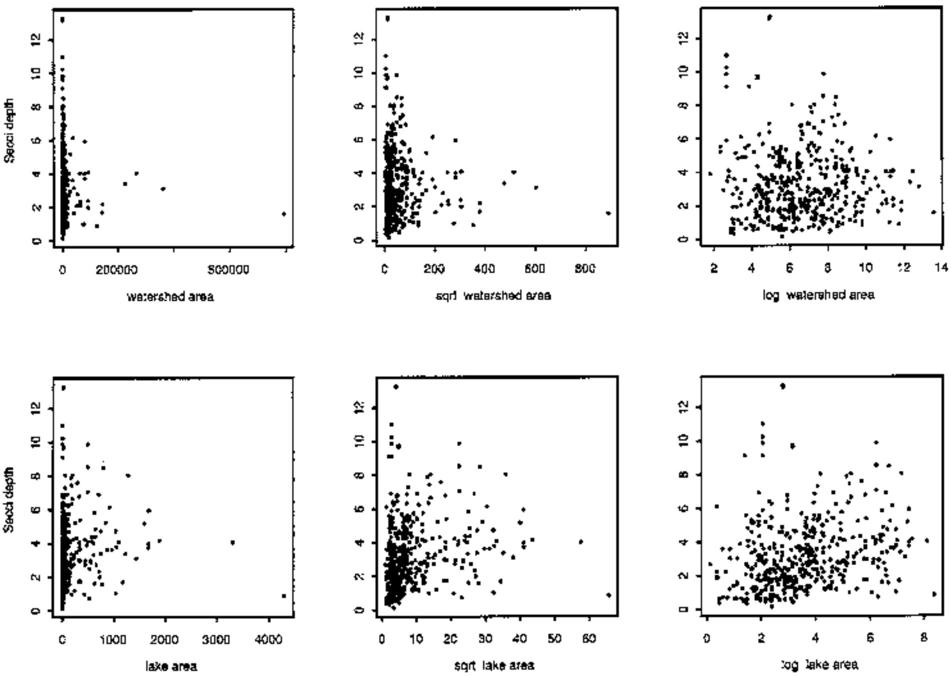

Figure 2. The auxiliary variable and response for real populations 12–17 (Table 2 and Section 3.1).

$\rho x_i + u_i$, where $-1 \leq \rho \leq 1$ is the desired correlation (between the π and the y) and u_i are independent and identically distributed normal random variables with mean 0 and variance $1 - \rho^2$. As with the auxiliary variable, the response is then standardized to have unit variance.

We created universes with 1,000 elements and used sample sizes of 50 and 100 elements. These sizes translate to sample proportions of 0.05 and 0.10. An estimate of the EMAP-lakes sample proportion is 0.092.

3.3 Response Structure for Simulations

The principal performance criteria we evaluated from the simulations are the estimators' empirical bias and mean square error (MSE). For an arbitrary estimator, S^2, the bias is defined as $E\left[S^2 - \sigma^2\right]$, where the expectation is approximated by the average of the estimates from the 10,000 simulated samples. The empirical MSE is defined as $E\left[\left(S^2 - \sigma^2\right)^2\right] = \text{bias}^2 + \text{var}(S^2)$ where $\text{var}(S^2)$ is approximated by the variance of the estimates from 10,000 simulated samples. An estimator's *relative* MSE (rMSE) with respect to the naive estimator's MSE is $E\left[\left(S^2 - \sigma^2\right)^2\right] / E\left[\left(s^2 - \sigma^2\right)^2\right]$.

For the artificial simulations, we simulated on a regular grid in the population space; then, we interpolated between the grid points creating a surface. We interpolated using the SPLUS (TM) function `interp`, which uses a method due to Akima (1978).

Table 3. The Estimators' Relative MSE (Section 3.3)

Pop.	*HT*	*HT-R*	*HT-GR*	*MOM*	*MOM-R*	*MOM-GR*	*YG*	*YG-R*	*YG-GR*
1	0.07	0.06	0.06	0.07	0.06	0.06	0.05	0.06	0.06
2	0	0	0	0	0	0	0	0	0
3	7.40	0.04	0.15	7.17	0.07	0.06	0.10	0.04	0.04
4	264.10	4.43	4.76	18.00	8.70	8.40	28.00	4.60	5.60
5	435.80	1.85	7.80	9.20	3.39	2.49	9.90	1.88	2.15
6	4590.50	21.60	125.40	386.60	102.90	89.40	514.30	22.50	40.20
7	7.40	7.00	6.80	10.10	9.40	9.30	12.30	7.40	7.90
8	1.51	1.49	1.49	1.60	1.58	1.58	1.67	1.55	1.56
9	7.70	3.78	3.28	18.30	10.90	10.00	24.30	3.95	5.70
10	2.59	2.50	2.46	3.10	2.96	2.93	3.62	2.60	2.69
11	2.01	1.98	1.97	2.29	2.23	2.21	2.52	2.05	2.09
12	6992.00	1.85	79.00	733.70	12.40	8.10	53.90	1.88	2.86
13	3.57	2.65	2.53	4.70	4.48	4.07	6.10	2.75	2.98
14	1.57	1.62	1.61	1.70	1.76	1.75	2.05	1.67	1.68
15	106.50	12.60	19.70	45.30	31.00	25.50	57.30	13.10	16.10
16	4.01	3.32	3.32	4.34	3.82	3.72	5.30	3.44	3.55
17	2.01	1.62	1.64	1.97	1.73	1.73	2.14	1.68	1.69

4. RESULTS

4.1 Results From the Simulations on Real Populations

The rMSE for the design-based estimators demonstrated the following general trends HT-R < HT-GR < HT, MOM-GR≈ MOM-R < MOM, and YG-R < YG-GR < YG (Table 3). These trends are dominated by the trend in the estimators' variances. Across the design-based estimators the general trend was HT-R ≈ YG-R < MOM-R.

For the two populations with high correlation between the auxiliary variable and the response (0.83 and 0.97 for populations 1 and 2, Table 2), the design-based estimators performed better (smaller MSE) than the naive estimator (Table 3). When the correlation was small (< 0.22, populations 4–17), however, the naive estimator performed better than the design-based estimators even in the cases of high inclusion probability variance (populations 6 and 12). For population 3, $\rho = 0.68$, the ratio and generalized ratio estimators performed better than the naive estimator but the nonratio forms of the HT and MOM estimators did not except for the YG estimator, which seemed to act more like the ratio estimators (Table 3).

Finally there was some evidence that inclusion probability variance and the correlation interact. Under extreme inclusion probability variance (Populations 6 and 12, $\gamma = 0.57$, Table 2) the design-based estimators performed their worse for $\rho = 0.13$ (population 6) where the smallest rMSE of the lot is 21.64 for the HT-R estimator, but they fared much better for $\rho = -0.04$ (population 12), where the rMSE for the HT-R estimator is 1.85.

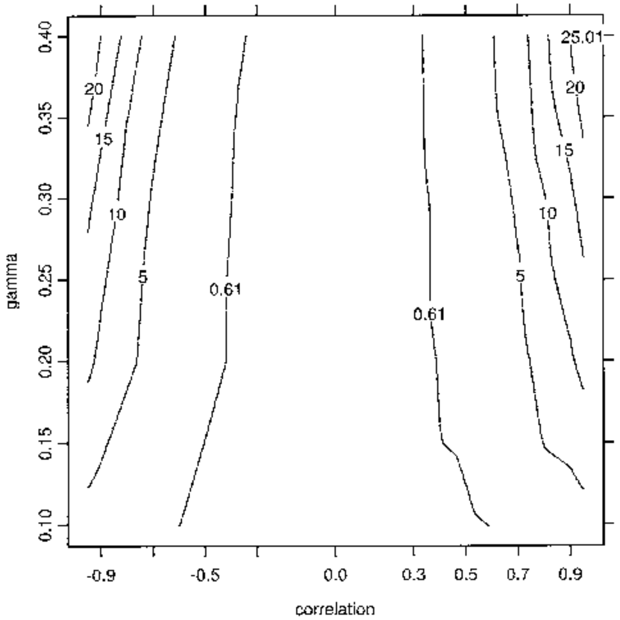

Figure 3. Empirical MSE for the naive estimator over the population space. The surface resembles a steep banked valley that runs along the inclusion probability variance direction and that thins out as γ increases, but only slightly.

4.2 Results From the Simulations on Artificial Populations

All estimators performed better as sample size was increased from 50 to 100 units; further, they improved at the same rate so that comparison of the estimators was invariant to sample size. The following discussion applies to both $n = 50$ and $n = 100$, but we present the results from only the $n = 50$ simulations.

The naive estimator performed well. Its MSE was up to six times smaller than the ratio estimators' MSEs when there was little correlation between the auxiliary and the response—approximately between ±0.4 at low levels of inclusion probability variance and ±0.3 at high levels of inclusion probability variance (Figures 3 and 4).

The unweighted design-based estimators, because of their large variances, performed terribly throughout the population space.

The performance of the three ratio estimators, as we saw in the real population simulations, depended on both population parameters. Under low inclusion probability variance, $\gamma < 0.25$, they performed best when the correlation ranged between 0.5 and 0.7; under high inclusion probability variance ($\gamma > 0.3$), on the other hand, they performed best when the correlation was between 0.3 and 0.5. They broke down when there was large inclusion probability variance and high correlation. Of these estimators, the HT-R estimator had the smallest MSE over most of the population space; however, the YG-R estimator's MSE was only three to six percent higher (Figure 5). The MOM-R estimator, on the other hand, performed worse than those two estimators; its MSE was up to eighty percent higher than that of the HT-R estimator (Figure 5). It fared worse however, when the correlation between inclusion probability and response was smallest, the region where we suggested using the naive estimator anyway (Figure 4), and its MSE was less than the other ratio estimators' MSE in the region of high correlation and low inclusion probability variance (Figure 5, lower corners).

The MOM-GR and the YG-GR estimators displayed slightly higher MSE than their ratio counterparts (Figure 5). The HT-GR estimator performed terribly throughout the population space due to its high variance.

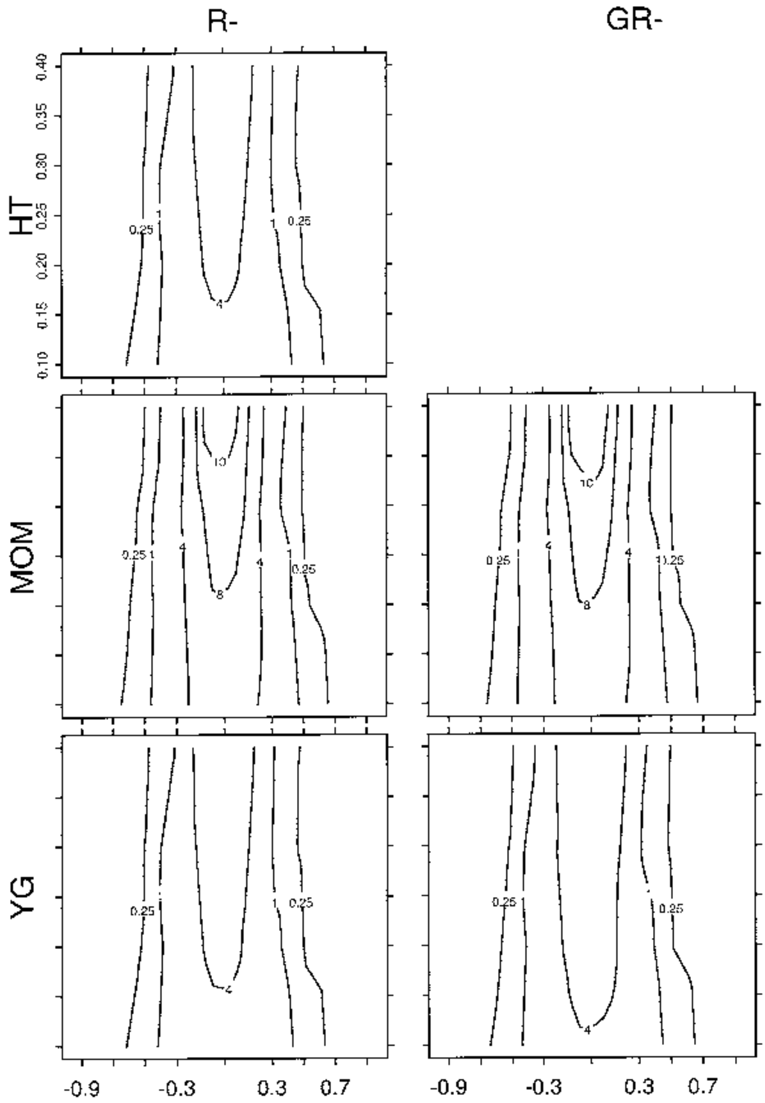

Figure 4. Relative mean square error (rMSE) for the ratio versions of the design-based estimators. The rMSE for the HT-GR estimator was larger than the rest and could not be displayed at this scale.

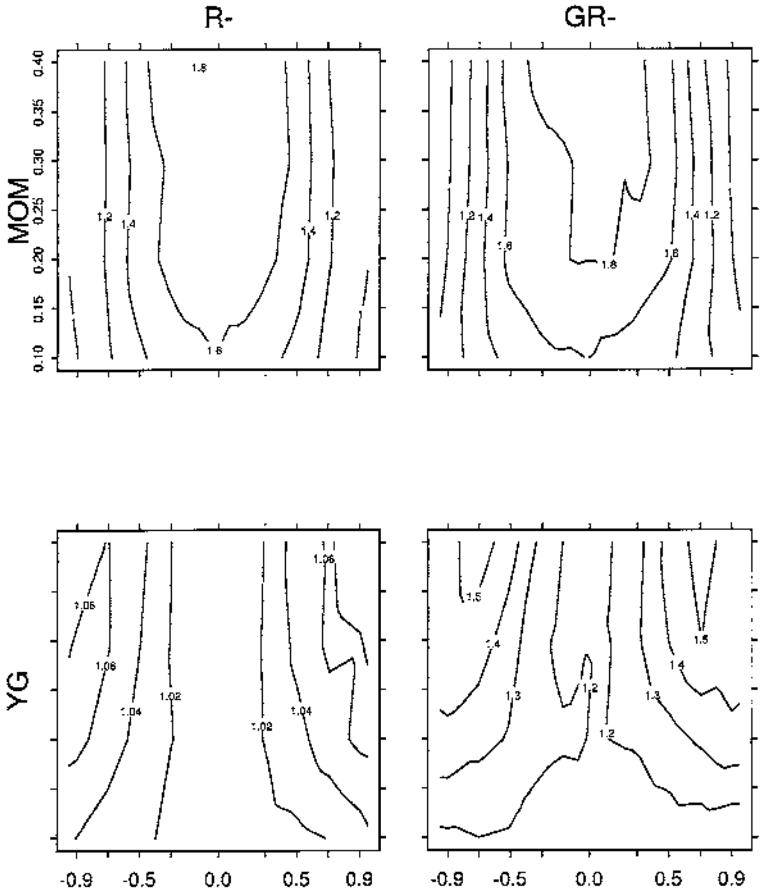

Figure 5. Ratio of empirical MSE for the MOM-R and YG-R estimators to the MSE of the HT-R estimator.

5. CONCLUSIONS

Our simulations suggest that the “naive” sample variance should work well despite being design biased, except when there is high correlation between the response and the auxiliary variable. Unfortunately, the sampling literature recommends that to estimate the population total one should find an auxiliary variable correlated with the response—this is the original motivation for πpx designs (Cochran 1977, chap. 10). In the surveys that interest us, however—multipurpose environmental surveys—high correlation is unlikely to occur for two reasons. First, an attempt to optimize the design for certain responses is likely to arrive at a poor design for other responses. Second, the inclusion probabilities are often based on practical or programmatic considerations rather than optimized for a specific response variable.

For other surveys, however, where πpx designs may be optimized for a particular response, we recommend using weighted estimators and, in particular, either the HT-R estimator or the YG-R estimator. The HT-R estimator has the added benefit that second order inclusion probabilities do not need to be calculated or approximated for estimation.

A final estimator, not included in this study, warrants mention because of its familiarity and ease of use. The statistical package SAS allows, in its general linear model procedure (`proc glm;`), the use of weighted estimates. It has been suggested that using these weights is a strategy for including sampling weights. The estimator that results is

$$S_{\mathrm{W}}^2 = \frac{1}{n-1}\left(\sum_{i=1}^{N} y_i^2 \frac{Z_i}{\pi_i} - \frac{1}{\hat{N}}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \frac{Z_i Z_j}{\pi_i \pi_j}\right).$$

Notice, this is approximately the HT-R estimator (Equation (2.2)) except for the $(n-1)$ divisor outside the sum and the $\frac{1}{\hat{N}}$ within the sum. As a result, if we naively plug our sample and inclusion probabilities into SAS we get a good estimator for the sum of square error but not the mean square error. The remedy is to use the correct "degrees of freedom," $(\hat{N}-1)$.

There are several caveats to this study; inferences are restricted to the designs we consider, approximations to the second-order inclusion probabilities, and the particular populations we examined. Our selection method (Section 3) provides an easy way to select a πps design that admits second-order inclusion probabilities. Other methods are more difficult to implement and/or restricted to particular situations $(n = 2)$ (Brewer and Hanif 1983; Cochran 1977). Although we feel that our results should apply in more general settings, we recommend that anyone who desires a πps design consider using this selection method. As for the approximations, calculation of the true inclusion probabilities is difficult so we feel most researchers will use the approximations in the end. Finally, our populations are simple; if a researcher has a population that does not resemble one considered here they should replicate this experiment for their situation; the authors are happy to share the code.

APPENDIX: THEOREMS AND PROOFS

This appendix provides the Horvitz-Thompson theorem and two corollaries used to develop the design-based estimators. Vector based notation allows for convenience in calculations and succinct notation (Dol, Steerneman, and Wansbeek 1996). Let $\vec{y}$, $\vec{\pi}$, and $\vec{Z}$ denote the $N \times 1$ vector of y_i's, π_i's, and Z_i's, and Π the $N \times N$ matrix of π_{ij}'s.

Define $\odot$ and $\oslash$ as elementwise multiplication and division, respectively (the Hadamard product). The symbols $\mathbf{1}$ and $\mathbf{I}$ denote an $N \times 1$ vector of ones and the $N \times N$ identity matrix, respectively.

To demonstrate vector notation, some population parameters are written in vector notation as follows. The mean is $\overline{y} = \frac{1}{N}\mathbf{1}'\vec{y}$ and the variance is $\sigma^2 = \frac{1}{(N-1)}\,\vec{y}'\left(\mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}'\right)\vec{y}$.

We review the Horvitz-Thompson (HT) theorem here and present two corollaries:

Theorem 1. *(Horvitz-Thompson) If unit $i \in \mathcal{U}$ is selected with probability π_i, where $\pi_i > 0$ for all $i \in \mathcal{U}$, then for any fixed $N \times 1$ vector λ, $\lambda'\left(\vec{y} \odot \vec{Z} \oslash \vec{\pi}\right)$ is unbiased for $\lambda'\vec{y}$.*

Proof: See Horvitz and Thompson (1952). □

Corollary 1. *Let the $N \times 1$ vector $\vec{z}$ be a second response on $\mathcal{U}$. Denote $\breve{Z} = \vec{Z}\vec{Z}' \oslash \Pi$. If $\pi_{ij} > 0$ for all $(i, j) \in \mathcal{U} \times \mathcal{U}$, then for any fixed $N \times N$ matrix* $\mathbf{A}$,

$$E\left[\vec{y}'\left(\mathbf{A} \odot \breve{Z}\right)\vec{z}\right] = \vec{y}'\mathbf{A}\vec{z}$$

.

Proof: $\vec{Z}$ is the only random element in the expectation and $E\left[\vec{Z}\vec{Z}'\right] = \Pi$ (Särndal et al. 1992). Quadratics result by taking $\vec{z} = \vec{y}$. □

Corollary 2. *Let z_{ij} be a response defined on $\mathcal{U} \times \mathcal{U}$. If $\pi_{ij} > 0$ for all $(i, j) \in \mathcal{U} \times \mathcal{U}$, then $\sum\sum_{ij \in s} z_{ij}/\pi_{ij}$ is unbiased for $\sum\sum_{ij \in U \times U} z_{ij}$*

Proof: See Särndal et. al. (1992, p. 48): Consider the universe $\mathcal{U} \times \mathcal{U}$, then the inclusion probabilities for "elements" will be the π_{ij} and the HT theorem applies. This approach to quadratic estimation is known as the batch approach in Liu and Thompson (1983). □

ACKNOWLEDGMENTS

Portions of this research were supported by the United States Environmental Protection Agency NNEMS grant number U914916-01-0 and by the National Science Foundation Vertical Integration in Research and Education (VIGRE) grant number DMS-9810726. This research was also partially funded by EPA cooperative agreements CR824682 and CR829096 to Oregon State University and CR829095 to Colorado State University. It has not been formally reviewed by EPA. The views expressed in this article are solely those of the authors; and EPA does not endorse any products or commercial services mentioned herein.

[Received November 2001. Revised September 2003.]

REFERENCES

Akima, H. (1978), "A Method of Bivariate Interpolation and Smooth Surface Fitting for Irregularly Distributed Data Points," *ACM Transactions on Mathematical Software*, 4, 148–159.

Brewer, K., and Hanif, M. (1983), *Sampling With Unequal Probabilities*, New York: Springer.

Cochran, W. G. (1977), *Sampling Techniques* (3rd Ed), New York: Wiley.

Dol, W., Steerneman, T., and Wansbeek, T. (1996), "Matrix Algebra and Sampling Theory: The Case of the Horvitz–Thompson Estimator," *Linear Algebra and its Applications*, 237, 225–238.

Hartley, H., and Rao, J. (1962), "Sampling With Unequal Probability and Without Replacement," *Annals of Mathematical Statistics*, 33, 350–374.

Hidiriglou, M., and Gray, G. (1980), "Construction of Joint Probability of Selection for Systematics p.p.s Sampling," *Applied Statistics*, 29, 107–112.

Horvitz, D., and Thompson, D. (1952), "A Generalization of Sampling Without Replacement from a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

Larsen, D., and Christie, S. (1993), "EMAP-Surface Waters 1991 Pilot Report, Technical Report EPA/620/R-93/003," Technical Report, U.S. Environmental Protection Agency Washington D.C.

Larsen, D., Thorton, K., Urquhart, N., and Paulsen, S. (1995), "The Role of Sample Surveys for Monitoring the Condition of the Nation's Lakes," *Environmental Monitoring and Assessment*, 32, 101–134.

Liu, T., and Thompson, M. (1983), "Properties of Estimators of Quadratic Finite Population Functions: The Batch Approach," *The Annals of Statistics*, 11, 275–285.

McDonald, T. (1996), "Analysis of Finite Population Surveys: Sample Size and Testing Considerations," Ph.D. thesis, Oregon State University.

Mitch, M., Kaufmann, P., Herlihy, A., Overton, W., and Sale, M. (1990), "National Stream Survey Database Guide. EPN600/8-90/055," Technical Report, U.S. EPA Environmental Research Laboratory, Corvallis; Oregon.

Royall, R., and Cumberland, W. (1981), "An Empirical Study of the Ratio Estimator and Estimators of its Variance," *Journal of the American Statistical Association*, 76, 66–88.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Stehman, S., and Overton, W. (1989), "Pairwise Inclusion Probability Formulas in Random-Order, Variable Probability, Systematic Sampling," Technical Report 131, Oregon State University Department of Statistics, Oregon State University, Corvallis, OR.

——— (1994a), "Comparison of Variance Estimators of the Horvitz–Thompson Estimator for Randomized Variable Probability Systematic Sampling," *Journal of the American Statistical Association*, 89, 30–43.

——— (1994b), "Environmental Sampling and Monitoring," in *Handbook of Statistics*, eds. G. Patil and C. Rao, Amsterdam: Elsevier Science.

Stevens, D. (1994), "Variable Density Grid-Based Sampling Designs for Continuous Spatial Populations," *Environmetrics*, 8, 167–195.

Thompson, S. K. (1992), *Sampling*, New York: Wiley.

Urquhart, N., Paulsen, S., and Larsen, D. (1998), "Monitoring for Policy-Relevant Regional Trends Over Time," *Ecological Applications*, 8, 246–257.

Yates, F., and Grundy, P. (1953), "Selection Without Replacement from Within Strata with Probability Proportional to Size," *Journal of the Royal Statistical Society*, Series B, 15, 253–261.